

Interface Reconciliation in Kahn Process Networks using CSP and SAT

Pavel Zaichenkov, Olga Tveretina, Alex Shafarenko

Compiler Technology and Computer Architecture Group,
University of Hertfordshire, United Kingdom
{p.zaichenkov,o.tveretina,a.shafarenko}@ctca.eu

Abstract. We present a new CSP- and SAT-based approach for coordinating interfaces of distributed stream-connected components provided as closed-source services. The Kahn Process Network (KPN) is taken as a formal model of computation and a Message Definition Language (MDL) is introduced to describe the format of messages communicated between the processes. MDL links input and output interfaces of a node to support flow inheritance and contextualisation. Since interfaces can also be linked by the existence of a data channel between them, the match is generally not only partial but also substantially nonlocal. The KPN communication graph thus becomes a graph of interlocked constraints to be satisfied by specific instances of the variables. We present an algorithm that solves the CSP by iterative approximation while generating an adjunct Boolean SAT problem on the way. We developed a solver in OCaml as well as tools that analyse the source code of KPN vertices to derive MDL terms and automatically modify the code by propagating type definitions back to the vertices after the CSP has been solved. Techniques and approaches are illustrated on a KPN implementing an image processing algorithm as a running example.

Keywords: coordination programming, component programming, Kahn Process Networks, interface coordination, constraint satisfaction, satisfiability

1 Introduction

The software intensive systems have reached unprecedented scale by every measure: number of lines of code; number of people involved in the development; number of dependencies between software components, and amount of data stored and manipulated [1]. Many of them include heterogeneous elements, which come from variety of different sources: parts of them are written in different languages and tuned for different hardware/software platforms. Furthermore, when the software is developed and modified by dispersed teams, inconsistencies in the design, implementation and usage are unavoidable. This leads to clashes of assumptions about operation cost, resource availability and algorithm processing rate. Last but not least, parts of the system are constantly changing. Many elements need to be replaced without negative effects on performance or behaviour of the rest of the system.

One way to attack the software challenge is to suggest a component-based design: a program is designed as a set of components, each represented by an interface that specifies how they can be used in an application, and one or more implementations which define their actual behaviour. When a designer of the application uses a component, they agree to rely only on the interface specification. Similarly, a developer who creates an implementation for a component is unaware of the context where the component will be used. An algorithm that specifies the behaviour depends solely on self-contained input and its result is produced in the form of a message without a specific destination address.

The process network, introduced by G. Kahn (KPNs) [2], is a collection of stream-connected algorithmic building blocks, which are fully independent single-threaded processes that lack a global state. The execution of the network generally requires a supervisory coordination program that manages the progress of the blocks and which provides a message-communication infrastructure for the streams. Since all domain-specific computations are performed by the sequential processes inside the blocks, programming is naturally separated into algorithm and concurrency engineering [3]. The coordination language is responsible for component orchestration, namely 1) dynamic load control and adaptivity for a changing environment; 2) access control to shared resources; and 3) communication safety between components. This paper focuses on the last aspect. Component-based design requires an implementation of a single component to be independent from the rest of the network. It raises a number of software engineering issues: components' interfaces are required to be specific enough so that components are aware of data structures communicated between them and, at the same time, generic enough to facilitate decontextualisation and software reuse.

In this paper we present a solution to the interface reconciliation problem for an interface definition language specifically designed for KPNs. We demonstrate a static mechanism (based on solving Constraint Satisfaction Problem (CSP) and SAT) that checks compatibility of component interfaces connected in a network with support of overloading and structural subtyping. This allows one to design completely decontextualised components, so that they may be reused in different contexts without changing the code. This is especially important when the components are provided as a compiled library and its source is either private or unavailable. The components are compatible with a potentially unlimited number of input/output data formats coming from the environment. We also introduce a *flow inheritance* [4] mechanism: put simply, a message sent from one component to another may also be required to contain additional data which, although not needed by the recipient itself, is nevertheless required by a component that the recipient sends its own messages to (Fig. 1).

We propose a Message Definition Language (MDL) that enables components' generic interfaces as well as subtyping and flow inheritance; we then recast the interface reconciliation problem as a CSP for the interface variables and propose an original solution algorithm that solves it by iterative approximation while generating an adjunct Boolean SAT problem on the way.

We designed and implemented a communication protocol¹ for components coded in C++ to demonstrate the capabilities of MDL. We developed tools that 1) automatically derive MDL interfaces from the source code; 2) generate a set of constraints given a netlist² that describes the topology of the network; 3) solve the CSP; and 4) based on the solution of the CSP generate compilable code for every component with some API provided for run-time support.

The process is similar to template specialisation in C++, however, in our case, constraints that are produced by a pair of vertices may affect the whole network, and, consequently, a global constraint satisfaction procedure is required. In this paper we provide a formal description of MDL, the CSP definition and the algorithm designed to solve the CSP.

Throughout the paper we demonstrate the utility of the proposed approach on a practical example: an image segmentation algorithm based on k-means clustering (Fig. 2).

Related work. Linda [5] is the first language designed to separate computation and coordination models. It is based on a simple tuplespace model. One of the disadvantages of the model is that the knowledge about the communication protocol is required while implementing the processes. The problem of separation of concerns has not been solved in Concurrent Collections from Intel [6] (Linda’s successor). Therefore, generic components, which may be reused in multiple contexts without being modified, are not supported in the language.

In the programming language Reo [7] components are communicating through hierarchical connectors that coordinate their activities and manipulate message dataflow. Similar to our approach, a constraint satisfaction engine, which finds a solution that specifies a valid interaction between components, is implemented. S. Kemper describes a SAT-based verification of Timed Constraint Automata that is used for coordination of communicating components [8] as well as in Reo. However, the research mostly focuses on the design of reusable interaction protocols and lacks the description of reusable component interfaces.

In previous years there were attempts to design efficient programming languages and run-time systems for parallel programming based on KPN [4,9], however, the interface reconciliation problem stemming from nonlocal inheritance in KPNs has not been given enough attention.

2 Motivation

Kahn Process Networks is a concurrency model that introduces data streams in the form of sequential channels that connect independent processes into a network. Decontextualisation of processes is an advantage of the model. Since processes do not share any data, a process’s conformity with the context is defined by its interface, which describes the kinds of message that the process

¹ for the avoidance of doubt we state that the term “protocol” is used here in the sense of ‘convention governing the structure and interpretation of messages’ and not in any state-transition sense

² a textual representation of a graph

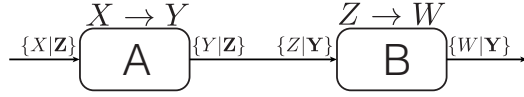


Fig. 1: Illustration of flow inheritance. A component **A** can process a value of type X and return a value of type Y as a result. However, an input message contains not only an element of type X , but also an element of type Z . The latter can be processed by a component **B**. Flow inheritance provides a mechanism for partial message processing in a pipelined fashion.

can send and receive. Our goal is to provide a means of interface coordination that supports genericity, i.e. the ability of an interface to function correctly in a variety of contexts.

The distributed components are commonly provided as closed-source services. Each service contains multiple processing functions compatible with a variety of contexts. The input interface of a component is specialised based on the message format that the message producer is capable to produce, and, symmetrically, the output interface is specialised based on the consumer's requirements. For example, one can define a component that contains two functions with type signatures $\text{Int} \rightarrow \text{Int}$ and $\text{Int} \rightarrow \text{String}$. The functions implement algorithms that compute different values given the same input. The task is to statically choose the algorithm based the consumer's requirements. This also demonstrates that input and output types in the interfaces of the KPN components are treated in the same manner.

The latter makes services fundamentally different from functions. In functional languages, a type signature that corresponds to the interface of the component in the example can be defined by the intersection type $\text{Int} \rightarrow \text{Int} \wedge \text{Int} \rightarrow \text{String}$, which is unsound due to its ambiguity. In functional languages, the return type of a function depends solely on input argument types. In contrast, the interfaces of the KPN components form context-dependent relations that offer a selection of output types to a consumer. This makes the typing decisions essentially nonlocal and genuinely multiple.

The problem being solved can be seen as a type inference problem; however, it cannot be solved using conventional type inference mechanisms based on first-order unification due to the presence of polymorphic output types and potential cyclic dependencies in the network (the example in Fig. 2 contains a back edge).

A common communication pattern in streaming networks is a pipeline, where a message travels along a chain of components that work on its content. The component can accept a subtype of the input type, but part of the message may be bypassed to another component down the pipeline if the message contains the data the further component will need to use (Fig. 1). Two modes of flow inheritance are envisaged, considered next.

Flow inheritance for records. The fundamental type of a message in a variety of systems is *record*, which is a collection of label-value pairs. Each component processes only a specific set of pairs, however the pairs that the component does not require may be bypassed to the output, so they can be processed in the

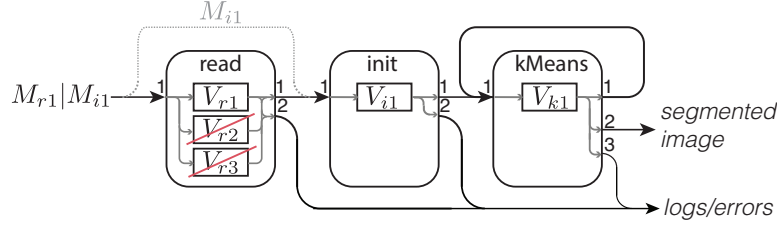


Fig. 2: An image segmentation algorithm based on k-means clustering that is implemented as a Kahn Process Network

next stages of the pipeline if they are required. For example, a message that represents a geometric shape and has a type $\{x:\text{float}, y:\text{float}, \text{radius}:\text{float}\}$ may be processed in two steps: the first component processes the position of the shape as defined by the pairs x and y , and the second one needs the pair labelled **radius**.

Flow inheritance for variants. OOP extensively uses overloading [10] to improve modularity and reusability of code. Similarly, support of polymorphic components in KPNs facilitates their reuse in different contexts. At the top level we see a component's interface as a collection of alternative label- record pairs, called *variants*, where the label corresponds to the particular implementation that can process the message defined by the given record, e.g.

$(: \text{cart}: \{x:\text{float}, y:\text{float}\}, \text{polar}: \{r:\text{float}, \text{phi}:\text{float}\} :).$

Here the colonised parentheses delimit the collection of variants and each variant is associated with a record written as a set of label-value pairs. Any message that does not belong to one of the accepted variants must cause an error unless there exists another component further down in the pipeline that can process it. In this case, the message should be bypassed to the recipient.

In streaming networks flow inheritance alleviates the problem and makes configuration of individual interfaces independent from each other. In our work we developed a solution for the interface reconciliation based on the CSP with support of flow inheritance for records and variants. Our mechanism statically detects implementation variants in components that are not required in the context, which is important for applications running in the Cloud where a user is charged proportionally to the amount of resources their application uses.

Example. As a running example we use our implementation [11] of an image segmentation algorithm based on k-means clustering [12]. The applications's KPN graph is shown in Fig. 2. The network represents a pipeline composed of three components:

- The component **read** opens an image file using an input message M_{r1} with the file name, and sends it to the first output channel. The component contains 3 functions that overload component's behaviour (i.e. the input interface of the component is defined by 3 variants): 1) V_{r1} loads the colour image

- in RGB format; 2) V_{r2} loads the greyscale image as an intensity one; and 3) V_{r3} loads the image as it is stored in the file.
- The component `init` sets initial parameters for the k-means algorithm. The component contains one processing function V_{i1} . The input message can either come from the component `read` or from the environment with an input message M_{i1} if it has been opened and preprocessed before. The input message must contain the number of clusters K and the image itself.
- The `kMeans` component represents an iterative implementation (defined as a function V_{k1}) of the k-means algorithm. The result of each iteration is sent to the first output channel, which is circuited back to the input channel of the component itself. This kind of design gives an opportunity to manage system load in the run-time and execute the next iteration only when sufficient resources are available. Once the cluster centres have converged, the algorithm yields the result to the second output channel.

Using flow inheritance for variants M_{i1} is routed directly to the `init` component bypassing a component `read`. Using flow inheritance for records a parameter K that is contained in M_{r1} is implicitly bypassed through `read` to `init`.

The interface reconciliation algorithm is capable of finding out that V_{r2} and V_{r3} are not used with the provided input, and functions containing the implementations will be excluded from the generated code.

3 Message Definition Language

Now we define the Message Definition Language (MDL) that describes component interfaces. Each component has its associated input and output interface terms. A *message* is a collection of data entities, defined by a corresponding collection of *terms* that can contain term variables, Boolean variables and Boolean expressions.

Each term is either atomic or a collection in its own right. Atomic terms are *symbols*, which are identifiers used to represent standard C++ types, such as `int` or `string`. To account for subtyping (including the kinds that are not present in C++) we include three categories of collections (see Fig. 3): *tuples* that demand exact match and thus admit no structural subtyping, *records* that are subtyped covariantly (a larger record is a subtype) and *choices* that are contravariantly subtyped using set inclusion (a smaller choice is a subtype). The intention of these terms is to represent

1. extensible data records [13,14], where additional named fields can be introduced without breaking the match between the producer and the consumer and where fields can also be inherited from input to output records by lowering the output type, which is always safe;
2. data-record variants, where generally more variants can be accepted by the consumer than the producer is aware of, and where such additional variants can be inherited from the output back to the input of the producer — hence contravariance — again by raising the input type, which is always safe also.

Term variables correspond to four categories of terms. However, for the correctness of the algorithm it is important to distinguish variables that represent choices from variables that represent other term categories (due to two kinds of subtyping defined by the seniority relation in Definition 1). We use an *up-coerced* term variable, e.g. $\uparrow a$, to represent a choice term and a *down-coerced* term variable, e.g. $\downarrow a$, to represent any other term, i.e. a symbol, a tuple or a record. Formally,

$$\langle \text{term variable} \rangle ::= \uparrow \text{identifier} \mid \downarrow \text{identifier}$$

We use symbol \square instead of \uparrow or \downarrow symbols in the context where the sort is unimportant, e.g. $\square a$ is a term variable that can be either up-coerced or down-coerced.

For brevity, term variables are called *variables*, Boolean variables are called *flags* and Boolean expressions are called *guards*. The following grammar specifies the guards:

$$\langle \text{bool} \rangle ::= (\langle \text{bool} \rangle \wedge \langle \text{bool} \rangle) \mid (\langle \text{bool} \rangle \vee \langle \text{bool} \rangle) \mid \neg \langle \text{bool} \rangle \mid \mathbf{true} \mid \mathbf{false} \mid \text{flag}$$

MDL terms are built recursively using the constructors: tuple, record, choice and switch, according to the following grammar:

$$\begin{aligned} \langle \text{term} \rangle &::= \langle \text{symbol} \rangle \mid \langle \text{term variable} \rangle \mid \langle \text{tuple} \rangle \mid \langle \text{record} \rangle \mid \langle \text{choice} \rangle \mid \langle \text{switch} \rangle \\ \langle \text{tuple} \rangle &::= (\langle \text{term} \rangle [\langle \text{term} \rangle]^*) \\ \langle \text{record} \rangle &::= \{ [\langle \text{label} \rangle (\langle \text{bool} \rangle) : \langle \text{term} \rangle, \langle \text{label} \rangle (\langle \text{bool} \rangle) : \langle \text{term} \rangle]^* [\mid \downarrow \text{identifier}]] \} \\ \langle \text{choice} \rangle &::= (: [\langle \text{label} \rangle (\langle \text{bool} \rangle) : \langle \text{term} \rangle, \langle \text{label} \rangle (\langle \text{bool} \rangle) : \langle \text{term} \rangle]^* [\mid \uparrow \text{identifier}]] :) \\ \langle \text{label} \rangle &::= \langle \text{symbol} \rangle \end{aligned}$$

Informally, a *tuple* is an ordered collection of terms and a *record* is an extensible, unordered collection of guarded labeled terms, where *labels* are arbitrary symbols, which are unique within a single record. A *choice* is a collection of alternative terms. The syntax of choice is the same as that of record except for the delimiters. The difference between records and choices is in subtyping and will become clear below when we define seniority on terms. We use choices to represent polymorphic messages and component interfaces.

Records and choices are defined in *tail form*. The tail is denoted by a variable that represents a term of the same kind as the construct in which it occurs. For example, in the term $\{l_1(\mathbf{true}): t_1, \dots, l_n(\mathbf{true}): t_n \mid \downarrow v\}$ the variable $\downarrow v$ represents the tail of the record, i.e. its members with labels $l_i : l_i \neq l_1, \dots, l_i \neq l_n$. A *switch* is a set of unlabeled (by contrast to a choice) guarded alternatives.

$$\langle \text{switch} \rangle ::= \langle \text{bool} \rangle : \langle \text{term} \rangle [, \langle \text{bool} \rangle : \langle \text{term} \rangle]^* \rangle$$

Exactly one guard must be **true** for any valid switch. The switch is substitutionally equivalent to the term marked by the **true** guard:

$$\langle \mathbf{false}: t_1, \dots, \mathbf{true}: t_i, \dots, \mathbf{false}: t_n \rangle = \langle \mathbf{true}: t_i \rangle = t_i.$$

The switch is an auxiliary construct intended for building conditional terms. For example, $\langle a: \text{int}, \neg a: \text{string} \rangle$ represents the symbol *int* if $a = \mathbf{true}$, and the symbol *string* otherwise.

For each term t , we use $\mathcal{V}^\uparrow(t)$ to denote the set of up-coerced term variables that occur in t , $\mathcal{V}^\downarrow(t)$ to denote the set of the down-coerced ones, and $\mathcal{F}(t)$ to denote the set of flags.

A term t is called *semi-ground* if it does not contain variables, i.e. $\mathcal{V}^\uparrow(t) \cup \mathcal{V}^\downarrow(t) = \emptyset$. A term t is called *ground* if it is semi-ground and does not contain flags, i.e. $\mathcal{V}^\uparrow(t) \cup \mathcal{V}^\downarrow(t) \cup \mathcal{F}(t) = \emptyset$.

A term t is *well-formed* if it is ground and one of the following holds:

1. t is a symbol.
2. $n > 0$ and t is a tuple $(t_1 \dots t_n)$ where all t_i , $0 < i \leq n$, are well-formed.
3. $n \geq 0$ and t is a record $\{l_1(b_1): t_1, \dots, l_n(b_n): t_n\}$ where for all $0 \leq i \neq j \leq n$, $b_i \wedge b_j \rightarrow l_i \neq l_j$ and all t_i for which b_i are **true** are well-formed.
4. $n \geq 0$ and t is a choice $(:l_1(b_1): t_1, \dots, l_n(b_n): t_n:)$ where for all $0 \leq i \neq j \leq n$, $b_i \wedge b_j \rightarrow l_i \neq l_j$ and all t_i for which b_i are **true** are well-formed.
5. $n > 0$ and t is a switch $\langle b_1: t_1, \dots, b_n: t_n \rangle$ where for some $1 \leq i \leq n$, $b_i = \mathbf{true}$ and t_i is well-formed and where $b_j = \mathbf{false}$ for all $j \neq i$.

If an element of a record, choice or switch has a guard that is equal to **false**, then the element can be omitted, e.g.

$$\{l_1(b_1): t_1, l_2(\mathbf{false}): t_2, l_3(b_n): t_3\} = \{l_1(b_1): t_1, l_3(b_n): t_3\}.$$

If an element of a record or a choice has a guard that is **true**, the guard can be syntactically omitted, e.g.

$$\{l_1(b_1): t_1, l_2(\mathbf{true}): t_2, l_3(b_n): t_3\} = \{l_1(b_1): t_1, l_2: t_2, l_3(b_n): t_3\}.$$

We define the *canonical form* of a well-formed collection as a representation that does not include **false** guards, and we omit **true** guards anyway. The canonical form of a switch is its (only) term with a **true** guard, hence any term in canonical form is switch-free.

Next we introduce a seniority relation on terms for the purpose of structural subtyping. In the sequel we use **nil** to denote the empty record $\{\}$, which has the meaning of **void** type in **C++** and represents a message without any data. Similarly, we use **none** to denote the empty choice $(: \cdot :)$.

Definition 1 (Seniority relation). *The seniority relation \sqsubseteq on well-formed terms is defined in canonical form as follows:*

1. **none** $\sqsubseteq t$ if t is a choice.
2. $t \sqsubseteq \mathbf{nil}$ if t is any term but a choice.
3. $t \sqsubseteq t$.
4. $t_1 \sqsubseteq t_2$, if for some $k, m > 0$ one of the following holds:
 - (a) $t_1 = (t_1^1 \dots t_1^k)$, $t_2 = (t_2^1 \dots t_2^m)$ and $t_1^i \sqsubseteq t_2^i$ for each $1 \leq i \leq k$;
 - (b) $t_1 = \{l_1^1: t_1^1, \dots, l_1^k: t_1^k\}$ and $t_2 = \{l_2^1: t_2^1, \dots, l_2^m: t_2^m\}$, where $k \geq m$ and for each $j \leq m$ there is $i \leq k$ such that $l_1^i = l_2^j$ and $t_1^i \sqsubseteq t_2^j$;
 - (c) $t_1 = (:l_1^1: t_1^1, \dots, l_1^k: t_1^k:)$ and $t_2 = (:l_2^1: t_2^1, \dots, l_2^m: t_2^m:)$, where $k \leq m$ and for each $i \leq k$ there is $j \leq m$ such that $l_1^i = l_2^j$ and $t_1^i \sqsubseteq t_2^j$;

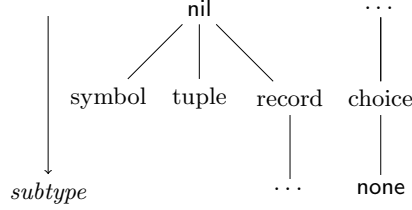


Fig. 3: Two semilattices representing the seniority relation for terms of different categories. The lower terms are the subtypes of the upper ones.

Given the relation $t \sqsubseteq t'$, we say that t' is *senior* to t and t is *junior* to t' .

Proposition 1. *The seniority relation \sqsubseteq is trivially a partial order, and (T, \sqsubseteq) is a pair of upper and lower semilattices (Fig. 3).*

The seniority relation represents the subtyping relation on terms. If a term t' describes the input interface of a component, then the component can process any message described by a term t , such that $t \sqsubseteq t'$.

Although the seniority relation is straightforwardly defined for ground terms, terms that are present in the interfaces of components can contain variables and flags. Finding such ground term values for the variables and such Boolean values for the flags that the seniority relation holds represents a CSP problem, which is formally introduced next.

4 Constraint Satisfaction Problem for KPN

In this section we define a Constraint Satisfaction Problem for Kahn Process Networks (CSP-KPN). We regard a KPN network as a directed weakly connected labeled graph $G = (V, E, L)$, where

1. V is a set of vertices. The vertices correspond to individual Kahn processes.
2. E is a set of edges, where each edge $e \in E$ is an ordered pair of vertices (v, v') , $v, v' \in V$. The edges correspond to channels connecting Kahn processes.
3. A function $L: E \rightarrow \text{Term} \times \text{Term}$ assigns a label³ to each edge $e \in E$ which represents a pair of MDL terms $L(e) = t \sqsubseteq t'$ called a *constraint*⁴. It defines the input requirements and the output properties associated with the channel.

Given a graph $G = (V, E, L)$ we define the set of constraints as

$$\mathcal{C}(G) = \bigcup_{e \in E} L(e),$$

³ we use the same word “label” to refer to the mark on a graph edge and the symbol that labels a term in a record or a choice; however our intention is always clear from the context.

⁴ in the rest of the paper symbol \sqsubseteq denotes the seniority relation for a pair ground terms; alternatively, if the terms are not ground, \sqsubseteq specifies a constraint.

the sets of up-coerced term variables $\mathcal{V}^\uparrow(\mathbf{G})$ and down-coerced term variables as $\mathcal{V}^\downarrow(\mathbf{G})$

$$\mathcal{V}^\uparrow(\mathbf{G}) = \bigcup_{t \sqsubseteq t' \in \mathcal{C}(\mathbf{G})} \mathcal{V}^\uparrow(t) \cup \mathcal{V}^\uparrow(t') \quad \text{and} \quad \mathcal{V}^\downarrow(\mathbf{G}) = \bigcup_{t \sqsubseteq t' \in \mathcal{C}(\mathbf{G})} \mathcal{V}^\downarrow(t) \cup \mathcal{V}^\downarrow(t'),$$

and the set of flags $\mathcal{F}(\mathbf{G})$ as

$$\mathcal{F}(\mathbf{G}) = \bigcup_{t \sqsubseteq t' \in \mathcal{C}(\mathbf{G})} \mathcal{F}(t) \cup \mathcal{F}(t').$$

Assume a vector of flags $\vec{f} = (f_1, \dots, f_l)$, a vector of term variables $\llbracket \vec{v} \rrbracket = (\llbracket v_1 \rrbracket, \dots, \llbracket v_m \rrbracket)$, a vector of Boolean values $\vec{b} = (b_1, \dots, b_l)$ and a vector of terms $\vec{s} = (s_1, \dots, s_m)$. Then for each term t

1. $t[\vec{f}/\vec{b}]$ denotes the vector obtained as a result of the simultaneous replacement of f_i with b_i for each $1 \leq j \leq l$;
2. $t[\llbracket \vec{v} \rrbracket/\vec{s}]$ denotes the vector obtained as a result of the simultaneous replacement of $\llbracket v_i \rrbracket$ with s_i for each $1 \leq i \leq m$;
3. $t[\vec{f}/\vec{b}, \llbracket \vec{v} \rrbracket/\vec{s}]$ is a shortcut for $t[\vec{f}/\vec{b}][\llbracket \vec{v} \rrbracket/\vec{s}]$.

Assume a KPN graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{L})$ such that $|\mathcal{F}(\mathbf{G})| = l$, $|\mathcal{V}^\uparrow(\mathbf{G})| = m$, $|\mathcal{V}^\downarrow(\mathbf{G})| = n$ and for some $l, m, n \geq 0$.

Definition 2 (CSP-KPN). We define a CSP for a KPN graph \mathbf{G} (CSP-KPN) as follows: for each $t \sqsubseteq t' \in \mathcal{C}(\mathbf{G})$ find a vector of Boolean values $\vec{b} = (b_1, \dots, b_l)$, a vectors of ground terms $\vec{t} = (t_1, \dots, t_m)$, $\vec{t}' = (t'_1, \dots, t'_n)$ such that

$$t[\vec{f}/\vec{b}, \uparrow \vec{v}/\vec{t}, \downarrow \vec{v}/\vec{t}'] \sqsubseteq t'[\vec{f}/\vec{b}, \uparrow \vec{v}/\vec{t}, \downarrow \vec{v}/\vec{t}'],$$

where $\vec{f} = (f_1, \dots, f_l)$, $\uparrow \vec{v} = (\uparrow v_1, \dots, \uparrow v_m)$, $\downarrow \vec{v} = (\downarrow v_1, \dots, \downarrow v_n)$. The tuple $(\vec{b}, \vec{t}, \vec{t}')$ is called a solution.

A CSP-KPN is decidable since the message definition language we introduced can be seen as a term algebra, and decision problems for term algebras are decidable [15].

5 Adjunct SAT

The CSP-KPN solution algorithm presented in the next section is iterative and takes advantage of the order-theoretical structure of the MDL (Proposition 1).

Let $\mathbf{B}_0 \subseteq \mathbf{B}_1 \subseteq \dots \subseteq \mathbf{B}_s$ be sets of Boolean constraints, and \vec{a}^\uparrow and \vec{a}^\downarrow be vectors of semiground terms such that $|\vec{a}^\uparrow| = |\mathcal{V}^\uparrow(\mathbf{G})|$ and $|\vec{a}^\downarrow| = |\mathcal{V}^\downarrow(\mathbf{G})|$. The vectors \vec{a}^\uparrow and \vec{a}^\downarrow are *conditional approximations* of the solution.

We seek the solution as a fixed point of a series of approximations in the following form:

$$(\mathbf{B}_0, \vec{a}_0^\uparrow, \vec{a}_0^\downarrow), \dots, (\mathbf{B}_{s-1}, \vec{a}_{s-1}^\uparrow, \vec{a}_{s-1}^\downarrow), (\mathbf{B}_s, \vec{a}_s^\uparrow, \vec{a}_s^\downarrow),$$

Algorithm 1 CSP-KPN(G)

```
1:  $c \leftarrow |\mathcal{C}(G)|$ 
2:  $i \leftarrow 0$ 
3:  $B_0 \leftarrow \emptyset$ 
4:  $\vec{a}_0^\uparrow \leftarrow (\text{none}, \dots, \text{none})$ 
5:  $\vec{a}_0^\downarrow \leftarrow (\text{nil}, \dots, \text{nil})$ 
6: repeat
7:   for  $1 \leq j \leq c : t_j \sqsubseteq t'_j \in \mathcal{C}(G)$  do
8:      $(B_{i \cdot c + j}, \vec{a}_{i \cdot c + j}^\uparrow, \vec{a}_{i \cdot c + j}^\downarrow) \leftarrow \text{SOLVE}(B_{i \cdot c + j - 1}, \vec{a}_{i \cdot c + j - 1}^\uparrow, \vec{a}_{i \cdot c + j - 1}^\downarrow, \text{true}, t_j, t'_j)$ 
9:   end for
10:   $i \leftarrow i + 1$ 
11: until  $(\text{SAT}(B_{i \cdot c}), \vec{a}_{i \cdot c}^\uparrow, \vec{a}_{i \cdot c}^\downarrow) = (\text{SAT}(B_{(i-1) \cdot c}), \vec{a}_{(i-1) \cdot c}^\uparrow, \vec{a}_{(i-1) \cdot c}^\downarrow)$ 
12: if  $B_{i \cdot c}$  is unsatisfiable then
13:   return Unsat
14: else
15:   return  $(\text{SATSol}(B_{i \cdot c}), \vec{a}_{i \cdot c}^\uparrow[\vec{f}/\vec{b}], \vec{a}_{i \cdot c}^\downarrow[\vec{f}/\vec{b}])$ 
16: end if
```

where for every $1 \leq k \leq s$ and a vector of Boolean values \vec{b} that is a solution to $\text{SAT}(B_k)$ (by $\text{SAT}(B_k)$ we mean a set of Boolean vector satisfying B_k):

$$\vec{a}_{k-1}^\uparrow[\vec{f}/\vec{b}] \sqsubseteq \vec{a}_k^\uparrow[\vec{f}/\vec{b}] \quad \text{and} \quad \vec{a}_k^\downarrow[\vec{f}/\vec{b}] \sqsubseteq \vec{a}_{k-1}^\downarrow[\vec{f}/\vec{b}], \quad (1)$$

where the elements of the vectors are compared pairwise. The starting point is $B_0 = \emptyset$, $\vec{a}_0^\uparrow = (\text{none}, \dots, \text{none})$, $\vec{a}_0^\downarrow = (\text{nil}, \dots, \text{nil})$ and the series terminates as soon as $\text{SAT}(B_s) = \text{SAT}(B_{s-1})$, $\vec{a}_s^\uparrow = \vec{a}_{s-1}^\uparrow$, $\vec{a}_s^\downarrow = \vec{a}_{s-1}^\downarrow$.

The adjunct set of Boolean constraints potentially expands at every iteration of the algorithm by inclusion of further logic formulas called *assertions* into its conjunction as the algorithm processes constraints $\mathcal{C}(G)$. Whether the set of Boolean constraints actually expands or not can be determined by checking the satisfiability of $\text{SAT}(B_k) \neq \text{SAT}(B_{k-1})$ for the current iteration k .

We argue below that if the original CSP-KPN is satisfiable then so is $\text{SAT}(B_s)$ and that the tuple of vectors $(\vec{b}_s, \vec{a}_s^\uparrow[\vec{f}/\vec{b}_s], \vec{a}_s^\downarrow[\vec{f}/\vec{b}_s])$ is a solution to the former, where \vec{b}_s is a solution of $\text{SAT}(B_s)$. In other words, the iterations terminate when the conditional approximation limits the term variables, and when the adjunct SAT constrains the flags enough to ensure the satisfaction of all CSP-KPN constraints. In general, the set $\text{SAT}(B_s)$ can have more than one solution. We select one of them, denoted by $\text{SATSol}(B_s)$ in the algorithm.

6 Algorithm

In this section we present Algorithm 1 which solves CSP-KPN for a given KPN graph $G = (V, E, L)$. It performs the following steps.

The algorithm iterates over the set of constraints $\mathcal{C}(G)$ and at each step it builds a closer approximation of the solution. The relation between two consequent approximations satisfies formulas (1).

The function $\text{SOLVE}()$ solves the constraint $t_j \sqsubseteq t'_j$ (see equation (2) in Lemma 1) and updates the vectors $\vec{a}_{i \cdot c+j}^\uparrow$ and $\vec{a}_{i \cdot c+j}^\downarrow$ with new values. Furthermore, it adds Boolean assertions presented below that ensure 1) satisfaction of the constraint for any $\vec{b} \in \text{SAT}(\mathbf{B}_{i \cdot c+j})$ as provided by Definition 1; and 2) well-formedness of the terms occurring in it.

The algorithm terminates if $\mathbf{B}_{i \cdot c} \equiv \mathbf{B}_{(i-1) \cdot c}$, $\vec{a}_{i \cdot c}^\uparrow = \vec{a}_{(i-1) \cdot c}^\uparrow$ and $\vec{a}_{i \cdot c}^\downarrow = \vec{a}_{(i-1) \cdot c}^\downarrow$.

Well-formedness assertions for records and choices. Any pair of elements in a well-formed record/choice cannot have equal labels. Therefore, for each record $\{l_1(b_1): t_1, \dots, l_1(b_n): t_n\}$ and each choice $(:l_1(b_1): t_1, \dots, l_1(b_n): t_n:)$ occurring anywhere in $\mathcal{C}(\mathbf{G})$ the following assertion is added to the SAT:

$$\bigwedge_{\forall 1 \leq i, j \leq n: l_i = l_j} \neg(b_i \wedge b_j).$$

Well-formedness assertions for switches. A well-formed switch term must have exactly one positive guard. Hence, for each switch $\langle b_1: t_1, \dots, b_n: t_n \rangle$ occurring anywhere in $\mathcal{C}(\mathbf{G})$ the following assertion is added to the SAT:

$$(b_1 \vee \dots \vee b_n) \wedge \bigwedge_{\forall 1 \leq i, j \leq n: i \neq j} \neg(b_i \wedge b_j).$$

Order assertions. We generate two kinds of order assertions.

1. If a variable $\llbracket x$ is junior to two incommensurable, identically guarded terms $\llbracket x \sqsubseteq \langle \dots : b : t_1 \dots \rangle$ and $\llbracket x \sqsubseteq \langle \dots : b : t_2 \dots \rangle$, where neither $t_1 \sqsubseteq t_2$ nor $t_2 \sqsubseteq t_1$, the assertion $\neg b$ is added to the adjunct SAT.
2. For each $c \in \mathcal{C}(\mathbf{G})$ of the form $\langle b_1: t_1, \dots, b_n: t_n \rangle \sqsubseteq \langle b'_1: t'_1, \dots, b'_m: t'_m \rangle$, the assertion $\neg(b_i \wedge b'_j)$ is added to the adjunct SAT if $t_i \not\sqsubseteq t'_j$.

Further details are found in Appendix A.

Lemma 1 (Loop invariant). *Algorithm 1 finds a series of approximations in the form of*

$$(\mathbf{B}_{k_0}, \vec{a}_{k_0}^\uparrow, \vec{a}_{k_0}^\downarrow), \dots, (\mathbf{B}_{k_{s-1}}, \vec{a}_{k_{s-1}}^\uparrow, \vec{a}_{k_{s-1}}^\downarrow), (\mathbf{B}_{k_s}, \vec{a}_{k_s}^\uparrow, \vec{a}_{k_s}^\downarrow),$$

where $k_i = i \cdot |\mathcal{C}(\mathbf{G})|$, and such that the following holds for any $\vec{b}_{k_i} \in \text{SAT}(\mathbf{B}_{k_i})$.

1. $\mathbf{B}_{k_i} \supseteq \mathbf{B}_{k_{i-1}}$;
2. $\vec{a}_{k_{i-1}}^\uparrow[\vec{f}/\vec{b}_{k_i}] \sqsubseteq \vec{a}_{k_i}^\uparrow[\vec{f}/\vec{b}_{k_i}]$ and $\vec{a}_{k_i}^\downarrow[\vec{f}/\vec{b}_{k_i}] \sqsubseteq \vec{a}_{k_{i-1}}^\downarrow[\vec{f}/\vec{b}_{k_i}]$;
3. $\nexists (\vec{a}', \vec{a}'')$:
 - (a) $\vec{a}_{k_{i-1}}^\uparrow[\vec{f}/\vec{b}_{k_i}] \sqsubseteq \vec{a}'[\vec{f}/\vec{b}_{k_i}]$, $\vec{a}'[\vec{f}/\vec{b}_{k_i}] \sqsubseteq \vec{a}_{k_i}^\uparrow[\vec{f}/\vec{b}_{k_i}]$;
 - (b) $\vec{a}_{k_i}^\downarrow[\vec{f}/\vec{b}_{k_i}] \sqsubseteq \vec{a}''[\vec{f}/\vec{b}_{k_i}]$, $\vec{a}''[\vec{f}/\vec{b}_{k_i}] \sqsubseteq \vec{a}_{k_{i-1}}^\downarrow[\vec{f}/\vec{b}_{k_i}]$.

Proof. Let $c = |\mathcal{C}(\mathbf{G})|$. To construct $(\mathbf{B}_{k_i}, \vec{a}_{k_i}^\uparrow, \vec{a}_{k_i}^\downarrow)$ given $(\mathbf{B}_{k_{i-1}}, \vec{a}_{k_{i-1}}^\uparrow, \vec{a}_{k_{i-1}}^\downarrow)$, the algorithm iteratively calls $\text{SOLVE}()$ function (see Appendix A).

$$(\mathbf{B}_r, \vec{a}_r^\uparrow, \vec{a}_r^\downarrow) \leftarrow \text{SOLVE}(\mathbf{B}_{r-1}, \vec{a}_{r-1}^\uparrow, \vec{a}_{r-1}^\downarrow, \text{true}, t_j, t'_j)$$

where $r = k_{i-1} + j$, $1 \leq j \leq c$ and $k_i = k_{i-1} + c$. For any $t_j \sqsubseteq t'_j$, $\text{SOLVE}()$ constructs the Boolean constraints B_r , and finds \vec{a}_r^\uparrow and \vec{a}_r^\downarrow by solving the equation

$$t[\vec{f}/\vec{b}_{r-1}, \uparrow\vec{v}/\vec{a}_{r-1}^\uparrow, \downarrow\vec{v}/\vec{a}_r^\downarrow] = t'[\vec{f}/\vec{b}_{r-1}, \uparrow\vec{v}/\vec{a}_r^\uparrow, \downarrow\vec{v}/\vec{a}_{r-1}^\downarrow], \quad (2)$$

1. $B_{k_i} \supseteq B_{k_{i-1}}$ by construction: $\text{SOLVE}()$ only adds new Boolean constraints to the existing set.
2. $\text{SOLVE}()$ iteratively constructs the local approximation for each constraint. The series of local approximations converges to the global approximation.
3. Proof by contradiction. Assume that $(\vec{a}_r^\uparrow, \vec{a}_r^\downarrow)$ is a solution of (2) and there exists another solution (\vec{a}', \vec{a}'') , such that $\vec{a}' \neq \vec{a}_r^\uparrow$ and $\vec{a}'' \neq \vec{a}_r^\downarrow$. Then

$$t[\vec{f}/\vec{b}_{r-1}, \uparrow\vec{v}/\vec{a}_{r-1}^\uparrow, \downarrow\vec{v}/\vec{a}'] = t'[\vec{f}/\vec{b}_{r-1}, \uparrow\vec{v}/\vec{a}', \downarrow\vec{v}/\vec{a}_{r-1}^\downarrow].$$

Two ground terms are equal only if they represent the same term, and, therefore, $\vec{a}' = \vec{a}_r^\uparrow$ and $\vec{a}'' = \vec{a}_r^\downarrow$, which contradicts the initial assumption. \square

Theorem 1 (Termination). *CSP-KPN(G) terminates after a finite number of steps for any KPN graph G.*

Proof. For a given graph G the number of flags, variables and labels for records and choices is bounded. There are two ways to produce new terms: either to add entries with new labels to records and choices, or to substitute subterms for terms.

1. The number of new terms constructed by adding new entries is bounded because the number of labels in a given G is finite.
2. The number of terms constructed by substituting subterms for other terms is bounded because a) the number of variables is finite (the algorithm does not generate new variables); b) after the variables have been instantiated, the category of the term cannot be changed, otherwise, the seniority relation would be violated.

It implies that for each $\uparrow v \in \mathcal{V}^\uparrow(\mathbf{G})$ there exists a ground term \bar{t} , such that $\uparrow v \sqsubseteq \bar{t}$, and for each $\downarrow v \in \mathcal{V}^\downarrow(\mathbf{G})$ there exists a ground term \underline{t} , such that $\underline{t} \sqsubseteq \downarrow v$. Providing that $|\text{SAT}(B_{k_i})| \leq |\text{SAT}(B_{k_{i-1}})|$, the algorithm terminates after a finite number of steps. \square

Theorem 2. *Assume a KPN graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{L})$. The set of constraints $\mathcal{C}(\mathbf{G})$ is inconsistent if and only if CSP-KPN(G) returns Unsat.*

Proof. As the initial approximation the algorithm selects the weakest approximation $(\emptyset, (\text{none}, \dots, \text{none}), (\text{nil}, \dots, \text{nil}))$, it follows from Lemma 1 that the algorithm iterates over all possible approximations in consecutive order starting from $(B_{k_0}, \vec{a}_{k_0}^\uparrow, \vec{a}_{k_0}^\downarrow)$. Therefore, the algorithm cannot skip a solution if one exists. By Theorem 1 the algorithm terminates after a finite number of steps. Hence, it returns Unsat only if and only if the set of constraints $\mathcal{C}(\mathbf{G})$ cannot be satisfied. \square

```

message _1_init(vector<vector<double> img);
message _2_error(string msg);
variant _1_read_color(string fname) { _1_init(...); ... _2_error(...); }
variant _1_read_grayscale(string fname) {...}
variant _1_read_unchanged(string fname) {...}

```

(a) The source code

```

IN
1: (: read_color(c): {fname: string | $_rc},
   read_grayscale(g): {fname: string | $_rg},
   read_unchanged(u): {fname: string | $_ru} | $^r :)
OUT
1: (: init(or c g u): {img: vector<vector<double>>, | $_ro1 } | $^r :)
2: (: error(or c g u): {msg: string | $_ro2 } :)
$_rc <= $_ro1; $_rg <= $_ro1; $_ru <= $_ro1;
$_rc <= $_ro2; $_rg <= $_ro2; $_ru <= $_ro2;

```

(b) The interface

Fig. 4: The source code and the interface of the component **read** of the image processing algorithm

7 Communication Protocol

In this section we demonstrate interfaces with flow inheritance and code customisation using the example from Section 2. The interfaces are defined as choice-of-records terms. Labels in the choice term of the input interface correspond to function names that can process messages tagged with corresponding labels. Output messages are produced by calling special functions called *salvos*. The name of a salvo corresponds to one of the labels in the output choice term. The compatibility of two components connected by a channel is defined by the seniority relation.

Consider the source code and the interface of the component **read** in Fig. 4. Integers that have been added as prefixes to functions in the code specify the channels that messages are received from and sent to. In the interfaces we use prefixes $\$^$ and $\$_$ before identifiers to denote up- and down-coerced variables, respectively.

A tail variable $\r in the interface enables flow inheritance for choices: variants from the input channel that cannot be processed by the component (i.e. all variants but **read_color**, **read_grayscale** or **read_unchanged**), are absorbed by $\r . Thus, the messages of type M_{r1} that contain the name of the image file are processed by the component and the messages of type M_{i1} are inherited to the output and forwarded directly to the component **init**.

Flow inheritance for records is realised by down-coerced variables $\$_rc$, $\$_rg$, $\$_ru$, $\$_ro1$, $\$_ro2$, and a set of auxiliary constraints. A record in the input message contains an entry with the label K, which a processing function does not expect. After solving the CSP, the entry is added to the tail variable $\$_ro1$,

because the solver deduces that the element with the label `K` is required by the component `init`.

Furthermore, we use flags `c`, `g` and `u` to exclude the code that is not used in the context. The guards in the output interface employ the joint set of flags from the input variants that can fire salvos specified in the output interface. In the example all three functions can fire `init` and `error` salvos; accordingly, the salvos' guards are $c \vee g \vee u$. The solver deduces that the variants `read_grayscale` and `read_unchanged` cannot receive any messages, and, therefore, their respective processing functions can be excluded from the code.

To facilitate decontextualisation we introduce a wrapper for every component called a *shell*: an auxiliary configuration file that provides facilities for renaming labels in output records and choices and changing the routing of output messages.

The source code and the interfaces for the other two components are available in the repository [11].

8 Implementation

We implemented the solver [16] for the CSP-KPN in `OCaml`. It works on top of the PicoSAT [17] library, the latter used as a subsolver dealing with Boolean assertions. The input for the solver is a set of constraints and the output is in the form of assignments to flags and term variables.

We also developed a toolchain in `C++` and `OCaml` that performs the interface reconciliation in five steps:

1. Given a set of `C++` sources (the components), augment them with macros acting as placeholders for the code that enables flow-inheritance.
2. Derive the interfaces from the code.
3. Given the interfaces and a netlist that specifies a KPN graph, construct the constraints to be passed on to the CSP-KPN algorithm.
4. Run the solver.
5. Based on the solution, generate header files for every component with macro definitions. In addition, the tool generates the API functions to be called when a component sends or receives a message.

Advantages of the presented design are the following:

- Interfaces and the code behind them can be generic as long as they are sufficiently configurable. No communication between component designers is necessary to ensure consistency in the design.
- Configuration and compilation of every component is separated from the rest of the application. This prevents source code leaks in proprietary software running in the Cloud.⁵

⁵ which is otherwise a serious problem. For example, proprietary `C++` libraries that use templates cannot be distributed in binary form due to restrictions of the language's static specialisation mechanism.

9 Conclusion and Future Work

We have presented a new static mechanism for coordinating component interfaces based on CSP and SAT that checks compatibility of component interfaces connected in a network with support of overloading and structural subtyping. We developed a fully decoupled Message Definition Language that can be used in the context of KPN for coordinating components written in any programming language. We defined the interface of C++ components to demonstrate the binding between the MDL and message processing functions. Our techniques support genericity, inheritance and structural subtyping, thanks to the order relation defined on MDL terms.

On the theory side, we presented the CSP solution algorithm, showed its correctness and identified the termination condition. Although we assume that the algorithm is NP-complete because of the SAT problem, which needs to be solved as a subproblem, the complexity of the algorithm will be evaluated in further research.

The next step will be to support multiple flow inheritance in the MDL, to enable combined structures with inheritance (for example, $(union \downarrow a \downarrow b)$ represents a record that contains a union of entries associated with records $\downarrow a$ and $\downarrow b$). This would allow one to design components that perform synchronisation and merge multiple messages into one while preserving the inheritance mechanism of a vertex.

In the context of Cloud, our results may prove useful to the software-as-a-service community since we can support much more generic interfaces than are currently available without exposing the source code of proprietary software behind them. Building KPNs the way we do could enable service providers to configure a solution for a network customer based on components that they have at their disposal as well as those provided by other providers and the customer themselves, all solely on the basis of interface definitions and automatic tuning to nonlocal requirements.

References

1. Northrop, L., Feiler, P., Gabriel, R.P., Goodenough, J., Linger, R., Longstaff, T., Kazman, R., Klein, M., Schmidtd, D., Sullivan, K., Wallnau, K.: Ultra-Large-Scale Systems: The Software Challenge of the Future. Technical report, Software Engineering Institute, Carnegie-Mellon (2006)
2. Kahn, G.: The Semantics of Simple Language for Parallel Programming. In: IFIP Congress. (1974) 471–475
3. Jesshope, C., Shafarenko, A.: Concurrency engineering. In: 13th Asia-Pacific Computer Systems Architecture Conference, ACSAC 2008. (2008) 1–8
4. Grelck, C., Scholz, S.B., Shafarenko, A.: A Gentle Introduction to S-Net: Typed Stream Processing and Declarative Coordination of Asynchronous Components. *Parallel Processing Letters* **18**(2) (2008) 221–237
5. Ahuja, S., Carriero, N., Gelernter, D.: Linda and Friends. *Computer* **19**(8) (August 1986) 26–34

6. Budimlić, Z., Burke, M., Cavé, V., Knobe, K., Lowney, G., Newton, R., Palsberg, J., Peixotto, D., Sarkar, V., Schlimbach, F., et al.: Concurrent Collections. *Scientific Programming* **18**(3) (2010) 203–217
7. Arbab, F.: A Channel-Based Coordination Model for Component Composition. In: *Mathematical Structures in Computer Science*, University Press (2002) 329–366
8. Kemper, S.: SAT-based Verification for Timed Component Connectors. *Science of Computer Programming* **77**(78) (2012) 779–798
9. Vrba, v., Halvorsen, P., Griwodz, C., Beskow, P., Espeland, H., Johansen, D.: The Nornir run-time system for parallel programs using Kahn process networks on multi-core machines — a flexible alternative to MapReduce. *The Journal of Supercomputing* **63**(1) (2013) 191–217
10. Strachey, C.: Fundamental concepts in programming languages. *Higher-order and symbolic computation* **13**(1-2) (2000) 11–49
11. Zaichenkov, P.: Image Segmentation using K-means Clustering. <https://github.com/zayac/joule/tree/kmeans> (June 2015) (accessed June 16, 2015).
12. MacQueen, J.: Some methods for classification and analysis of multivariate observations (1967)
13. Gaster, B.R., Jones, M.P.: A Polymorphic Type System for Extensible Records and Variants. Technical report (1996)
14. Leijen, D.: Extensible records with scoped labels (September 2005)
15. Ferrante, J., Rackoff, C.W.: The computational complexity of logical theories. *Lecture notes in mathematics*. Springer-Verlag (1979)
16. Zaichenkov, P.: The toolchain and the solver for interface reconciliation in KPNs. <https://github.com/zayac/joule> (June 2015) (accessed June 16, 2015).
17. Biere, A.: Picosat essentials. *Journal on Satisfiability, Boolean Modeling and Computation (JSAT)* (2008)

A Appendix: Solve function

Algorithm 2 SOLVE($B_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t, t'$)

```

1:  $\hat{B}_i \leftarrow \text{ASSERTWELLFORMED}(\text{ASSERTWELLFORMED}(B_i, b, t), b, t')$ 
2: if  $((t = \text{none} \text{ or } t' = \text{nil}) \text{ and } (t \neq \text{none} \text{ or } t' \neq \text{nil}))$ 
3:   or  $(t \text{ and } t' \text{ are ground, and } t = t')$  then
4:   return  $(\hat{B}_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow)$ 
5: else if  $t = \uparrow v \text{ and } t' = \uparrow v'$  then
6:    $B_{i+1}, \vec{a}_{i+1}^\uparrow \leftarrow$  set a new approximation for  $\uparrow v'$  equal to the one of  $\uparrow v$ 
7: else if  $t = \downarrow v \text{ and } t' = \downarrow v'$  then
8:    $B_{i+1}, \vec{a}_{i+1}^\downarrow \leftarrow$  set a new approximation for  $\downarrow v$  equal to the one of  $\downarrow v'$ 
9:   return  $(B_{i+1}, \vec{a}_i^\uparrow, \vec{a}_{i+1}^\downarrow)$ 
10: else if  $t$  is nil, symbol, tuple or record, and  $t' = \downarrow v'$  then
11:   return SOLVE( $\hat{B}_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t, t'[\uparrow \vec{v}/\vec{a}_i^\uparrow, \downarrow \vec{v}/\vec{a}_i^\downarrow]$ )
12: else if  $t$  is a choice and  $t' = \uparrow v'$  then
13:    $B_{i+1}, \vec{a}_{i+1}^\uparrow \leftarrow$  set a new approximation for  $\uparrow v'$  as  $t[\uparrow \vec{v}/\vec{a}_i^\uparrow, \downarrow \vec{v}/\vec{a}_i^\downarrow]$  when  $b$ 
14:   return  $(B_{i+1}, \vec{a}_{i+1}^\uparrow, \vec{a}_i^\downarrow)$ 
15: else if  $t = \downarrow v$ , and  $t'$  is nil, symbol, tuple or record then
16:    $B_{i+1}, \vec{a}_{i+1}^\downarrow \leftarrow$  set a new approximation for  $\downarrow v$  as  $t'[\uparrow \vec{v}/\vec{a}_i^\uparrow, \downarrow \vec{v}/\vec{a}_i^\downarrow]$  when  $b$ 
17:   return  $(B_{i+1}, \vec{a}_i^\uparrow, \vec{a}_{i+1}^\downarrow)$ 
18: else if  $t = \uparrow v$  and  $t'$  is a choice then
19:   return SOLVE( $\hat{B}_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t[\uparrow \vec{v}/\vec{a}_i^\uparrow, \downarrow \vec{v}/\vec{a}_i^\downarrow], t'$ )
20: else if  $t$  and  $t'$  are tuples then
21:   return SOLVETUPLE( $\hat{B}_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t, t'$ )
22: else if  $t = \text{nil}$  and  $t'$  is a record then
23:   return SOLVENILRECORD( $\hat{B}_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t'$ )
24: else if  $t$  and  $t'$  are records then
25:   return SOLVERECORDRECORD( $\hat{B}_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t, t'$ )
26: else if  $t$  and  $t'$  are choices then
27:   return SOLVECHOICECHOICE( $\hat{B}_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t, t'$ )
28: else if  $t$  or  $t'$  is a switch then
29:   return SOLVESWITCH( $\hat{B}_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t, t'$ )
30: else
31:   return  $(\hat{B}_i \cup \{-b\}, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow)$ 
32: end if

```

Algorithm 3 ASSERTWELLFORMED(B_i, b, t)

```
1: if  $t$  is a record  $\{l_1(b_1): t_1, \dots, l_n(b_n): t_n\}$  or  $(:l_1(b_1): t_1, \dots, l_n(b_n): t_n:)$  then
2:    $B_{i+1} \leftarrow B_i \bigcup_{\forall 1 \leq i, j \leq n: l_i = l_j} \{b \rightarrow \neg(b_i \wedge b_j)\}$ 
3: else if  $t$  is a switch  $\langle b_1: t_1, \dots, b_n: t_n \rangle$  then
4:    $B_{i+1} \leftarrow B_i \cup \{b_1 \vee \dots \vee b_n\} \bigcup_{\forall 1 \leq i, j \leq n: i \neq j} \{b \rightarrow \neg(b_i \wedge b_j)\}$ 
5: end if
6: return  $B_{i+1}$ 
```

Algorithm 4 SOLVETUPLE(TUPLE($B_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t, t'$))

```
1: Let  $t$  be of the form  $(t_1 \dots t_n)$ 
2:  $g \leftarrow t'[\uparrow \vec{v}/\vec{a}_i^\uparrow, \downarrow \vec{v}/\vec{a}_i^\downarrow]$ 
3: if  $g = (t'_1 \dots t'_m)$  and  $n = m$  then
4:   for  $i: 1 \leq j \leq n$  do
5:      $B_{i+j}, \vec{a}_{i+j}^\uparrow, \vec{a}_{i+j}^\downarrow \leftarrow \text{SOLVE}(B_{i+j-1}, \vec{a}_{i+j-1}^\uparrow, \vec{a}_{i+j-1}^\downarrow, b, t_j, t'_j)$ 
6:   end for
7:   return  $(B_{i+n}, \vec{a}_{i+n}^\uparrow, \vec{a}_{i+n}^\downarrow)$ 
8: else
9:   return  $(B_i \cup \{-b\}, \vec{a}_{i+n}^\uparrow, \vec{a}_{i+n}^\downarrow)$ 
10: end if
```

Algorithm 5 SOLVENILRECORD($B_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t'$)

```
1:  $g \leftarrow t'[\uparrow \vec{v}/\vec{a}_i^\uparrow, \downarrow \vec{v}/\vec{a}_i^\downarrow]$ 
2: Let  $g$  be of the form  $\{l'_1(b'_1): t'_1, \dots, l'_m(b'_m): t'_m\}$ 
3: return  $(B_i \cup \{b \rightarrow \bigwedge_{j=1}^n \neg b'_j\}, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow)$ 
```

Algorithm 6 SOLVERECORDRECORD($B_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t, t'$)

```
1:  $g \leftarrow t'[\uparrow \vec{v}/\vec{a}_i^\uparrow, \downarrow \vec{v}/\vec{a}_i^\downarrow]$ 
2: if  $t = \{l_1(b_1): t_1, \dots, l_n(b_n): t_n\}$  then
3:   for  $j: 1 \leq j \leq m$  do
4:     if  $\exists k: l_k \in t, l_k = l'_j$  then
5:        $B_{i+j}, \vec{a}_{i+j}^\uparrow, \vec{a}_{i+j}^\downarrow \leftarrow \text{SOLVE}(B_{i+j-1}, \vec{a}_{i+j-1}^\uparrow, \vec{a}_{i+j-1}^\downarrow, b \rightarrow b'_j \rightarrow b_k, t_k, t'_j)$ 
6:     else
7:        $B_{i+j} \leftarrow B_{i+j-1} \cup \{b \rightarrow \neg b'_j\}$ 
8:     end if
9:   end for
10: else if  $t = \{l_1(b_1): t_1, \dots, l_n(b_n): t_n \mid \downarrow v\}$  then
11:   for  $j: 1 \leq j \leq m$  do
12:     if  $\exists k: l_k \in t, l_k = l'_j$  then
13:        $B_{i+j}, \vec{a}_{i+j}^\uparrow, \vec{a}_{i+j}^\downarrow \leftarrow \text{SOLVE}(B_{i+j-1}, \vec{a}_{i+j-1}^\uparrow, \vec{a}_{i+j-1}^\downarrow, b \rightarrow b'_j \rightarrow b_k, t_k, t'_j)$ 
14:     else
15:        $B_{i+j}, \vec{a}_{i+j}^\uparrow, \vec{a}_{i+j}^\downarrow \leftarrow$  set a new approximation for  $\downarrow v$  as  $\{l'_j(b'_j): t'_j\}$  when  $b$ 
16:     end if
17:   end for
18: end if
19: return  $(B_{i+m}, \vec{a}_{i+m}^\uparrow, \vec{a}_{i+m}^\downarrow)$ 
```

Algorithm 7 SOLVECHOICECHOICE($B_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t, t'$)

```

1:  $g \leftarrow t[\uparrow \vec{v} / \vec{a}_i^\uparrow, \downarrow \vec{v} / \vec{a}_i^\downarrow]$ 
2: if  $t' = (:l_1(b_1): t_1, \dots, l_n(b_n): t_n:)$  then
3:   for  $j: 1 \leq j \leq m$  do
4:     if  $\exists k: l_k \in t, l'_k = l_j$  then
5:        $B_{i+j}, \vec{a}_{i+j}^\uparrow, \vec{a}_{i+j}^\downarrow \leftarrow \text{SOLVE}(B_{i+j-1}, \vec{a}_{i+j-1}^\uparrow, \vec{a}_{i+j-1}^\downarrow, b \rightarrow b_j \rightarrow b'_k, t_j, t'_k)$ 
6:     else
7:        $B_{i+j} \leftarrow B_{i+j-1} \cup \{b \rightarrow \neg b'_j\}$ 
8:     end if
9:   end for
10: else if  $t' = (:l_1(b_1): t_1, \dots, l_n(b_n): t_n \mid \uparrow v:)$  then
11:   for  $j: 1 \leq j \leq m$  do
12:     if  $\exists k: l_k \in t, l'_k = l'_j$  then
13:        $B_{i+j}, \vec{a}_{i+j}^\uparrow, \vec{a}_{i+j}^\downarrow \leftarrow \text{SOLVE}(B_{i+j-1}, \vec{a}_{i+j-1}^\uparrow, \vec{a}_{i+j-1}^\downarrow, b \rightarrow b_j \rightarrow b'_k, t_j, t'_k)$ 
14:     else
15:        $B_{i+j}, \vec{a}_{i+j}^\uparrow, \vec{a}_{i+j}^\downarrow \leftarrow$  set a new approximation for  $\uparrow v$  as  $(:l_j(b_j): t_j:)$  when  $b$ 
16:     end if
17:   end for
18: end if
19: return  $(B_{i+m}, \vec{a}_{i+m}^\uparrow, \vec{a}_{i+m}^\downarrow)$ 

```

Algorithm 8 SOLVESWITCH($B_i, \vec{a}_i^\uparrow, \vec{a}_i^\downarrow, b, t, t'$)

```

1: if  $t = \langle b_1: t_1, \dots, b_n: t_n \rangle$  then
2:   for  $j: 1 \leq i \leq n$  do
3:      $B_{i+j}, \vec{a}_{i+j}^\uparrow, \vec{a}_{i+j}^\downarrow \leftarrow \text{SOLVE}(B_{i+j-1}, \vec{a}_{i+j-1}^\uparrow, \vec{a}_{i+j-1}^\downarrow, b, t_j, t')$ 
4:   end for
5: else if  $t' = \langle b'_1: t'_1, \dots, b'_n: t'_n \rangle$  then
6:   for  $i: 1 \leq j \leq n$  do
7:      $B_{i+j}, \vec{a}_{i+j}^\uparrow, \vec{a}_{i+j}^\downarrow \leftarrow \text{SOLVE}(B_{i+j-1}, \vec{a}_{i+j-1}^\uparrow, \vec{a}_{i+j-1}^\downarrow, b, t, t'_j)$ 
8:   end for
9: end if
10: return  $(B_{i+n}, \vec{a}_{i+n}^\uparrow, \vec{a}_{i+n}^\downarrow)$ 

```
